CENTRO PAULA SOUZA
GOVERNO DO ESTADO DE SÃO PAULO

Faculdade de Tecnologia de Americana
Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas

APLICAÇÃO DO ALGORITMO *EXPECTATION MAXIMIZATION* PARA CLUSTERIZAÇÃO DE DADOS SOBRE MORTALIDADE INFANTIL

FABRICIO SCHMIDT GALEGO

Americana, SP
2013

CENTRO PAULA SOUZA
GOVERNO DO ESTADO DE SÃO PAULO

Faculdade de Tecnologia de Americana
Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas

APLICAÇÃO DO ALGORITMO *EXPECTATION MAXIMIZATION* PARA CLUSTERIZAÇÃO DE DADOS SOBRE MORTALIDADE INFANTIL

FABRICIO SCHMIDT GALEGO
fabricio.sgalego@gmail.com

Trabalho Monográfico, desenvolvido em cumprimento à exigência curricular do Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas da Fatec-Americana, sob orientação do Prof. Me. Kléber de Oliveira Andrade.

Área: Banco de Dados, Estatística

Americana, SP
2013

BANCA EXAMINADORA

Professor: Me. Kléber de Oliveira Andrade (Orientador)

Professor: Me. Maria Elizete Luz Saes

Professor: Rogério Nunes de Freitas

AGRADECIMENTOS

Agradeço primeiramente a Deus, pois com Sua permissão fui capaz de chegar até aqui.

Aos meus pais pela paciência e suporte.

Aos professores, por todas orientações ao longo do curso.

DEDICATÓRIA

Dedico este trabalho aos meus pais, pelo apoio e paciência durante toda minha jornada até aqui, e por terem proporcionado todo o suporte que me possibilitou lutar pelos meus objetivos.

“[...], mesmo que nos separemos por um breve período, temos que nos lembrar de que Deus é BOM e até mesmo as coisas mais TERRÍVEIS acontecem pela melhor das RAZÕES, [...]”

(MILLAR, MARK: Os Supremos, 2002)

RESUMO

Mineração de dados é o processo de obtenção de conhecimento em bases de dados, para transformá-los em informações úteis para a tomada de decisão ou avaliação de resultados. Com o avanço tecnológico, o número de informações produzidas no setor da saúde aumentou, porém este número de dados é difícil de ser interpretado para apoiar tomadas de decisões referentes ao planejamento de novos projetos de redução da taxa de mortalidade infantil no Brasil. O objetivo do presente estudo foi, portanto, exemplificar as técnicas de mineração de dados e implementar o algoritmo *Expectation-Maximization* em uma base de dados sobre óbitos infantis, procurando encontrar padrões que possam ser úteis. Entre as informações resultantes da aplicação deste algoritmo, observou-se que o maior grupo de óbitos infantis atinge significativamente os dados referentes à raça e ao tipo de parto. Apesar da análise dos *clusters* ser bastante complexa, informações como esta pode ser utilizada em futuros projetos e campanhas.

Palavras Chave: Mineração de Dados; Mortalidade Infantil; *Expectation-Maximization;*

ABSTRACT

Data mining is the process of obtaining knowledge in databases, to transform them into useful information for decision making or evaluation of results. With the technological advancement, the amount of information produced in the health sector increased, however this amount of data is difficult to be interpreted to support decision making regarding the planning of new projects to reduce the infant mortality rate in Brazil. The aim of this study was exemplify the data mining techniques and implement the Expectation Maximization algorithm on a database of infant deaths, trying to find patterns that might be useful. Among the information resulting from the application of this algorithm, it was observed that the largest group of infant deaths reaches significantly the data relating to race and type of childbirth. Although the analysis of clusters be quite complex, as this information can be used in future projects and campaigns.

Keywords: Data Mining; Infant Mortality; Expectation-Maximization.

LISTA DE SIGLAS

API	*Application Programming Interface* (*Interface* de Programação de Aplicações)
DATASUS	Departamento de Informática do SUS
EM	*Expectation Maximization* (Maximização da Expectância)
IBGE	Instituto Brasileiro de Geografia e Estatística
JVM	*Java Virtual Machine* (MáquinaVirtual Java)
KDD	*Knowledge-Discovery in Databases* (Descoberta de Conhecimento em Base de Dados)
OMS	Organização Mundial da Saúde
SIM	Sistema de Informação sobre Mortalidade
SUS	Sistema Único de Saúde
UCI	Universidade da Califórnia, Irvine
WEKA	*Waikato Enviroment for Knowledge Analysis*

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

INTRODUÇÃO

Atualmente o uso da informática é essencial para diversas tarefas do ser humano, e o contínuo avanço tecnológico faz com que novas informações sejam criadas a todo instante (BRAGA, 2005).

As unidades de saúde geralmente possuem um enorme conjunto de dados sobre seus pacientes, que vão desde informações pessoais como nome e endereço, até dados específicos como data da última consulta ou eventuais tratamentos. Porém é desvantajoso ter uma grande base de dados sem que esta lhe proporcione algum retorno, portanto o ideal seria interpretar tais dados de modo a transformá-los em informações significativas para tomada de decisão e desta forma trazer melhorias para o hospital e para a população que o utiliza. Uma das tecnologias utilizadas para se obter essas informações é o KDD (*Knowledge Discovery in Databases* - Descoberta de Conhecimento em Bases de Dados), sendo a mineração de dados (*Data Mining*) a parte mais importante deste procedimento (OLIVEIRA, 2001).

Mineração de dados é o processo de obtenção de conhecimento em bases de dados, que procura padrões entre os dados para transforma-los em informações úteis para a tomada de decisão ou avaliação de resultados (BRAGA, 2005).

O autor diz ainda que a mineração de dados possui várias técnicas, sendo mais utilizados Clusterização, que reúne os dados em grupos de acordo com similiariedades, Redes Neurais, que avaliam os dados através de nós interligados em uma rede, Árvore de Classificação, que testa todos os valores do dado para identificar aqueles que possuem mais influência nos itens de saída selecionados para exame, e Regressão, que estima a probabilidade de uma variável alvo com base no demais dados.

De acordo com Frank e Asuncion (2010) a mineração de dados pode ser aplicada em inúmeras situações práticas, como calcular a probabilidade de um jogador ganhar uma partida de jogo da velha com base em suas jogadas iniciais, ou até mesmo fazer previsão do clima e tempo. Os autores disponibilizaram no site da UCI (Universidade da Califórnia, Irvine) um repositório de bases de dados públicos para serem utilizados em estudos, além de diversos exemplos de aplicação.

Para tanto o estudo se justificou pelo fato de que os grandes hospitais atualmente possuem uma enorme base de dados de seus pacientes e as técnicas de mineração de dados auxiliam a transformação destes dados em informações úteis e fundamentais para tomada de decisão eficiente (OLIVEIRA, 2001).

O pesquisador pretende especializar-se na área de banco de dados e tratamento de informações, pois se interessa em seguir tal área como carreira profissional por considerar o assunto atrativo e primordial no desenvolvimento de softwares e na informática em geral.

Já o problema foi: O avanço tecnológico aumentou o número de informações produzidas no setor da saúde, incluindo dados a respeito da mortalidade infantil, porém esse grande número de dados é difícil de ser interpretado, sendo que, se fosse, poderia produzir conhecimento que ajudariam no planejamento de novos projetos com o objetivo de minimizar tal mau em nosso país.

Como pergunta que se buscou responder: A técnica de clusterização da mineração de dados pode extrair padrões suficientes em uma base de dados dos óbitos infantis a ponto de apontar grupos de dados com similariedades que podem auxiliar tomadas de decisão referentes à novos projetos que objetivam diminuir a taxa de mortalidade infantil no Brasil?

As hipóteses foram: a) O uso da técnica de clusterização é capaz de descobrir conhecimento que ajude o planejamento de projetos para redução da mortalidade infantil; b) Apesar de descobrir alguns padrões, o conhecimento resultante da aplicação do algoritmo *Expectation Maximization* é insuficiente para auxiliar a criação de novos projetos de redução da mortalidade infantil, e c) O processo de mineração de dados aplicado não é capaz de gerar conhecimento suficiente para diminuir a taxa de mortalidade infantil.

O objetivo geral consistiu em exemplificar as técnicas de mineração de dados, focando no funcionamento e nas etapas do algoritmo *Expectation Maximization*, procurando demonstrar a importância de seu uso para produzir uma compreensão dos dados que possa influenciar tomadas de decisões.

Os objetivos específicos foram: a) Fazer um levantamento bibliográfico sobre mortalidade infantil para apresentar ao leitor os conceitos básicos; b) Fazer um levantamento bibliográfico sobre mineração de dados, visando compreender

adequadamente a técnica para apresentar os conceitos básicos e a história da mineração de dados; e c) Aplicar o algoritmo *Expectation Maximization* sobre uma base de dados dos óbitos infantis disponibilizados pelo Ministério da Saúde. Através da análise desta aplicação espera-se produzir conhecimento suficiente para otimizar a compreenção dos dados e auxiliar a criação de futuros projetos de minimização da taxa de mortalidade infantil no Brasil.

Como metodologia para o desenvolvimento deste trabalho, foi utilizada pesquisa aplicada, que Andrade (2009) conceitua como a pesquisa que põe em prática os conceitos estudados gerando conhecimento para solucionar um problema específico. Esta pesquisa é adequada, pois será desenvolvido um modulo de implementação de mineração de dados a partir de estudos desta tecnologia. Na forma de abordagem foi utilizada a pesquisa quantitativa, que para Andrade (2009), consiste em extrair conhecimento de dados numéricos e estatísticos, o que se enquadra neste trabalho, pois o modulo implementado tratará os dados apresentando diversos novos dados estatísticos que serão analisados para avaliar possíveis influência de fatores sobre a mortalidade infantil no Brasil.

Quanto aos objetivos, foi empregada a metodologia exploratória, que Andrade (2009) define como a pesquisa que harmoniza os aspectos teóricos e práticos, o que será feito com o desenvolvimento do modelo proposto a fim de estimular a compreensão. Foi utilizada também a metodologia descritiva, que segundo Andrade (2009), expõe as características de determinado assunto. Neste trabalho esta metodologia foi adotada para explicar os conceitos de Mineração de Dados, e da mortalidade infantil.

Como procedimentos técnicos, foi aplicada a pesquisa bibliográfica, que o autor conceitua como a pesquisa feita a partir de materiais já publicados, como livros e artigos. Esta pesquisa foi essencial para tornar possível conceituar e apresentar as técnicas utilizadas nos projetos de Mineração de Dados, e situar o leitor a respeito da mortalidade infantil.

O trabalho foi estruturado em quatro capítulos, sendo que o primeiro introduz o leitor ao assunto de mortalidade infantil expondo seu histórico tanto no Brasil como no exterior, o segundo traz um histórico da Mineração de Dados e seus conceitos, bem como suas técnicas de implementação, o terceiro descreve a preparação dos

dados e a aplicação do algoritmo proposto. Com base nas informações conseguidas a partir dos estudos realizados no capítulo anterior, o quarto capítulo se reserva às Considerações Finais.

1. INDICADORES DE SAÚDE

Este capítulo introduz o leitor ao assunto de mortalidade infantil, que é um dos indicadores de saúde utilizados para avaliar os serviços oferecidos à população e qualidade de vida da sociedade em questão.

A taxa de mortalidade infantil faz parte dos indicadores de mortalidade, que registram dados de diversas mortalidades específicas, como taxa de mortalidade por diabete melito, taxa de mortalidade por doenças transmissíveis e mortalidade neonatal. Além destes indicadores de mortalidade, o ministério da saúde também monitoram indicadores demográficos, socioeconômicos, de morbidade, de fatores de risco e proteção, de recursos e de cobertura.

Os indicadores demográficos são aqueles que monitoram os dados sobre o avanço da população, como grau de urbanização, esperança de vida, taxa de crescimento, entre outros. Os indicadores socioeconômicos monitoram as informações específicas da economia, como taxa de desemprego, taxa de trabalho infantil, PIB (produto interno bruto), e informações sociais, como a taxa de analfabetismo e escolaridade da população. Já os indicadores de morbilidade são específicos da área da saúde e registram informações a respeito da incidência e taxa de diversas doenças transmissíveis, taxas de internação hospitalar e prevalência de pacientes em diálise. Também na área da saúde, os indicadores de fatores de risco e de proteção armazenam dados sobre fatores que influenciam a saúde da população, como prevalência de fumantes, prevalência de excesso de peso e proporção de nascidos vivos de baixo peso ao nascer. Os indicadores de recursos monitoram dados de recursos disponíveis à população, como número de profissionais da saúde por habitante, número de leitos hospitalares por habitantes, e também informações a respeito de investimentos na área da saúde. Por fim, os indicadores de cobertura monitoram a abrangência dos serviços oferecidos, como cobertura vacinal, cobertura de coleta de lixo e cobertura de consultas de pré-natal.

Todos estes indicadores são imprescindíveis para avaliação da condição de vida na região em questão e são utilizados no planejamento de estratégias que objetivam aumentar o bem estar e a qualidade de vida da população.

1.1. TAXA DE MORTALIDADE INFANTIL

A taxa de mortalidade infantil equivale ao número de crianças menores de um ano que morrem entre mil nascidas vivas. Este número pode ser calculado em diferentes regiões, como cidades, estados, países, ou qualquer outro local, e é extremamente importante para que se possa avaliar a qualidade de vida do local em questão, uma vez que esta taxa é fortemente ligada aos serviços oferecidos à população, como disponibilidade de vacinas, sistema de saúde e saneamento básico (FRANCISCO, 2013).

Tabela 1 - Mortalidade Infantil em diversas regiões (FONTE: Adaptado de WHO, 2013).

Mortalidade Infantil			
	1990	**2000**	**2011**
Brasil	49	31	14
África	106	93	68
Américas	33	23	13
Sudeste da Ásia	77	59	42
Europa	27	17	11
Mediterrâneo Oriental	73	58	44
Pacífico Ocidental	37	27	13

De acordo com a Tabela 1, com dados extraídos do *World Health Statistics*, a taxa de mortalidade infantil no Brasil em 1990 era de 49 e caiu para 14 em 2011. Este número ainda é maior que a taxa calculada nas Américas, cuja taxa caiu de 33 para 13 no mesmo período, e também da Europa, que reduziu sua taxa de 27 para 11.

Nota-se também que o continente africano possui uma taxa de mortalidade infantil extremamente superior às das outras regiões, mostrando uma provável falta de estrutura no sistema de saúde local.

1.2. HISTÓRICO DA MORTALIDADE INFANTIL NO BRASIL

Assim como nas outras regiões ao redor do mundo, a taxa de mortalidade infantil também tem diminuído ao longo dos últimos anos. De acordo com os dados do IBGE (Instituto Brasileiro de Geografia e Estatística) de 1999, mostrados na Tabela 2, a taxa de mortalidade infantil no Brasil em 1930 era de 162,4, decrescendo para 48,3 em 1990. É importante salientar que segundo o IBGE (1999), mesmo em 1930 com esse valor sendo tão alto quando comparado aos valores mais recentes, esta taxa já estava passando por um processo de minimização com otimização dos serviços oferecidos à população.

Tabela 2 - Taxa de mortalidade infantil, segundo as Grandes Regiões Brasil - 1930/1990 (FONTE: adaptado de IBGE, 1999).

Ano	Taxa de mortalidade infantil (‰)					
	Brasil	Norte	Nordeste	Sudeste	Sul	Centro-Oeste
1930	162,4	193,3	193,2	153	121	146
1940	150	166	187	140	118	133
1950	135	145,4	175	122	109	119
1960	124	122,9	164,1	110	96	115
1970	115	104,3	146,4	96,2	81,9	89,7
1980	82,8	79,4	117,6	57	58,9	69,6
1990	48,3	44,6	47,3	33,6	27,4	31,2

Como pode-se notar na Tabela 2 e também no gráfico ilustrado na Figura 1, a taxa de mortalidade tem diminuido significativamente. Nota-se também que os valores entre as regiões no passado eram mais próximos, entretanto a diferença aumentou com o passar do tempo, principalmente entre as regiões sul, que em 1990 tinha taxa de 27,4, e nordeste, que tinha taxa de 47,3 no mesmo período. Desta forma, apesar da queda da taxa, ainda pode se perceber uma grande desigualdade entre tais regiões.

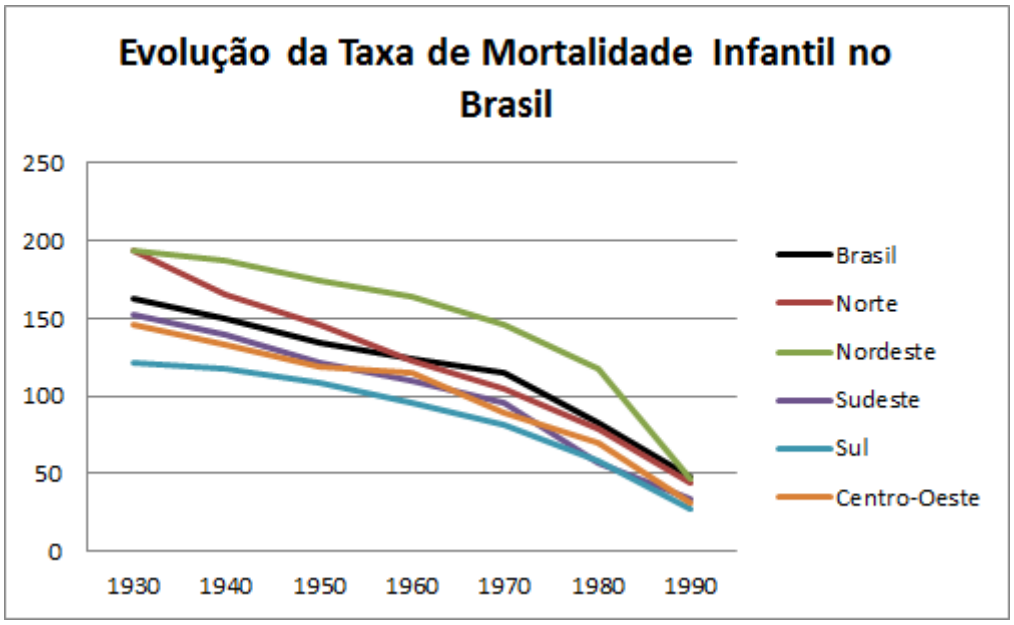


Figura 1 - Gráfico da taxa de mortalidade infantil nas regiões do Brasil (FONTE: IBGE, 1999).

Com base em tais dados, também é possível notar que a queda da taxa de mortalidade infantil acentuou-se fortemente a partir da década de 70. Segundo o IBGE (1999), isto se deve ao modelo de intervenção que foi implantado no saneamento básico, na medicina curativa, preventiva, e nos últimos anos na criação de programas de saúde que acompanham a gestante e o bebê desde o pré-natal, e campanhas de aleitamento materno e vacinação. Entretanto, este declínio não se deve somente ao investimento nos serviços oferecidos para a população, pois deve ser somada a queda acentuada do nível de fecundidade, porém este fator não tira a importância da qualidade dos serviços públicos, principalmente os investimentos no setor da saúde em áreas carentes, na evolução da taxa de mortalidade infantil no Brasil.

2. MINERAÇÃO DE DADOS

A mineração de dados surgiu no fim da década de 1980. Nesta época os bancos de dados relacionais tinham sido recentemente desenvolvidos e já eram implementados com sucesso, o que resultou em um enorme volume de dados (WILLIAMS, 2011).

No início dos anos 1990s, este aumento de dados produzidos fez com que os pesquisadores de Banco de Dados se preocupassem com tão grande número de dados armazenados e pouco utilizados. Isto fez com que estes pesquisadores se interessassem pela mineração de dados, para que estes dados pudessem produzir conhecimento útil para as empresas.

Porém esta tecnologia também atraiu profissionais de outras áreas, principalmente de Inteligência Artificial, que tem como um dos principais fundamentos a capacidade de aprendizado de máquinas através da análise de um conjunto de dados, o que se assemelhava com o objetivo da mineração de dados. A partir de então as duas linhas de pesquisas começaram a progredir em harmonia.

Outra área de pesquisa que contribuiu para o desenvolvimento da mineração de dados é a Estatística, muito utilizada na verificação da veracidade dos modelos, pois por meio de estatísticas podemos descobrir a relevância das relações entre os dados, mesmo que tais relações não estejam explícitas.

Atualmente, mineração de dados é uma tecnologia que chama cada vez mais a atenção de pesquisadores e empresas por ter se mostrado eficaz na análise de dados e extração de conhecimento.

2.1. DESCOBERTA DE CONHECIMENTO

O processo de transformar os enormes volumes de dados em informação útil é denominado Descoberta de Conhecimento em Bases de Dados, ou KDD (*Knowledge Discovery in Databases*). Esse processo é composto por várias etapas,

como mostrado na Figura 2, tendo início na seleção da base de dados e englobando a preparação e tratamento destes dados, a implementação da mineração de dados e, por fim, a análise dos resultados para descoberta de conhecimento (OLIVEIRA, 2001).

Após selecionar a base de dados, antes de aplicar as técnicas de mineração de dados é preciso preparar tais dados, selecionando os atributos que possam ter relevância na mineração e contribuir para a obtenção de conhecimento. Os dados estritamente específicos como nome ou telefone, ou dados constantes, devem ser descartados.

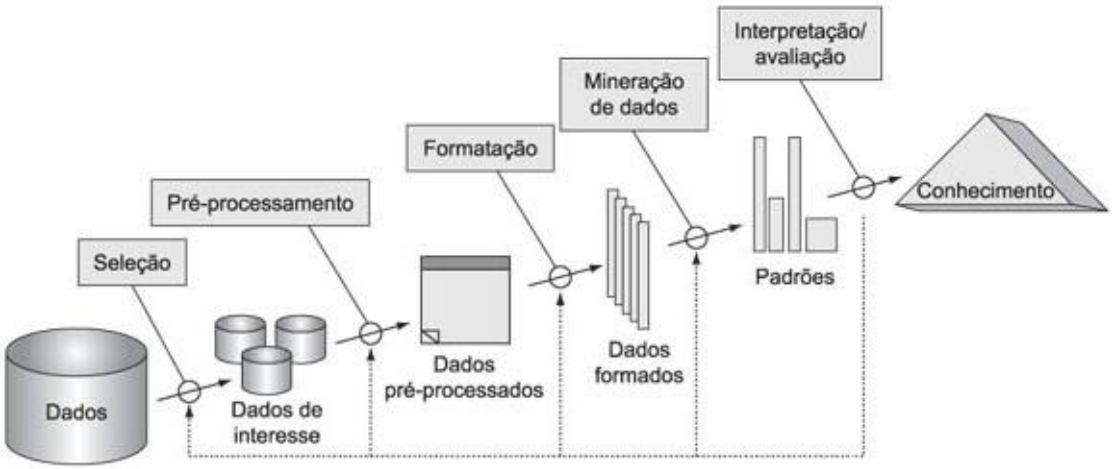


Figura 2 - Etapas do KDD (Fonte: FAYYAD et. al., 1996).

Na primeira etapa deve-se selecionar a base de dados na qual se deseja aplicar a mineração. É essencial que a base de dados contenha dados corretamente inseridos, com o menor número de erros de digitação, ausência de informações, dados aberrantes e ausência de valores, pois influenciarão na incoerência dos resultados. Por este motivo é importante que os dados passem pela próxima etapa, onde estes problemas serão corrigidos.

A segunda etapa consiste no pré-processamento dos dados. Nesta etapa os dados que não influenciam o objetivo do modelo que será construído deverão ser retirados. Para os dados inconsistentes ou ausentes são empregadas diversas técnicas para substituição destes valores, como, no caso de valores quantitativos,

preencher com a média de toda a base de dados, a média local dos dados próximos, ou a mediana (BRAGA, 2005).

Após a verificação da consistência dos dados, a base de dados deve ser convertida para um formato que possa ser utilizado pela aplicação onde será realizada a mineração. Outro objetivo desta etapa é reduzir o número de dados através de diversas técnicas, como a Sumarização que pode agrupar dados substituindo-os por intervalos, ou a Codificação, que consiste em substituir valores qualitativos em quantitativos para facilitar os cálculos estatísticos (BRAGA, 2005).

Com a base de dados pronta, a próxima etapa é a mineração dos dados, sendo considerada a principal etapa do KDD. Neste momento são aplicados um ou mais métodos específicos para extração de regras ou padrões, que depois de interpretados resultarão em conhecimento. É comum nesta fase voltar às etapas anteriores para fazer novas modificações nos dados.

Depois de aplicar a mineração, deve-se por em prática a última etapa, onde as regras e os padrões encontrados na etapa anterior serão analisados e convertidos em conhecimento e, desta forma, auxiliar a tomada de decisão para o alcance do objetivo, como por exemplo a otimização do Retorno do Investimento ou potencialização dos lucros (OLIVEIRA, 2001).

2.2. CONCEITUALIZAÇÃO

Segundo Fayyad et. al (1996, p 6):

> "[...] Extração de Conhecimento de Base de Dados é o processo de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis embutidos nos dados."

Porém, de acordo com Rezende et. al. (2003), alguns pesquisadores consideram o KDD como o processo completo da descoberta de conhecimento em bases de dados, sendo a mineração de dados a principal etapa.

Esta etapa do KDD envolve a aplicação repetida de métodos e algoritmos específicos para encontrar padrões nas bases de dados. Estes padrões serão interpretados na etapa final do KDD e proporcionará a descoberta de conhecimento (OLIVEIRA, 2001).

2.3. TAREFAS

A mineração de dados possui diversos métodos para modelagem, que devem ser escolhidos de acordo com a necessidade do modelo a ser construído. É importante relatar que não há um método específico para um modelo, e aconselha-se a implementação de mais de um algoritmo para minimizar as incertezas e escolher o que mais condiz com seus objetivos. Após a seleção do método e da implementação o modelo deve ser validado para que os resultados possam ser considerados admissíveis (BRAGA, 2005).

De acordo com Braga (2005), as tarefas para mineração de dados estão divididos em dois grandes grupos:

a) Predição, que estima o valor de uma variável com base nos demais atributos;

b) Descrição, o qual faz agrupamentos e correlações entre os atributos de uma base dados, para proporcionar maior compreensão das informações.

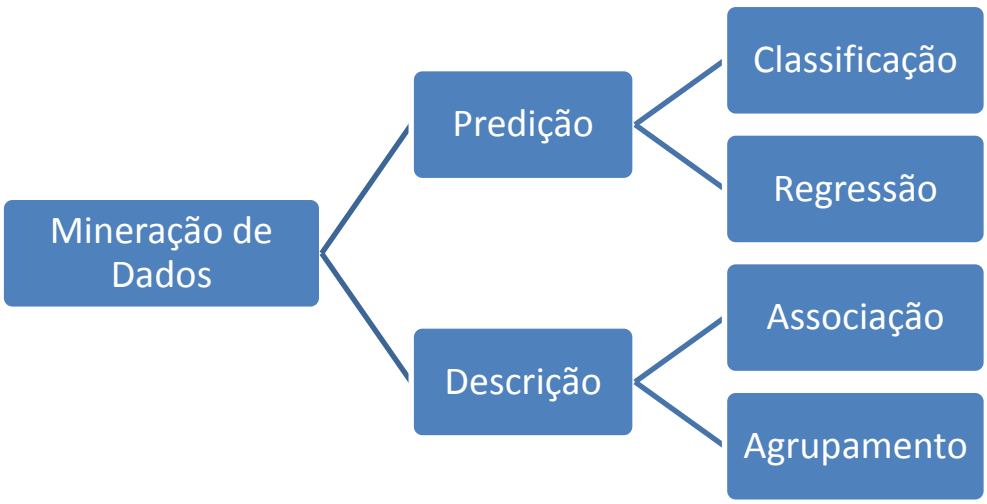


Figura 3 – Técnicas de mineração de dados (Fonte: Elaboração própria).

O primeiro grupo enquadra os métodos de Regressão e de Classificação, e no segundo se encontram os métodos de Associação e de Agrupamento, conforme mostrado na Figura 3. Cada um destes métodos, por sua vez, compreendem várias técnicas e algoritmos diferentes.

Os métodos de mineração de dados são divididos também em Aprendizado Supervisionado e Aprendizado Não-Supervisionado, sendo o primeiro quando a variável que se quer estimar é conhecida na base de dados e a mineração categoriza os demais dados em relação à ela, e a segunda quando a variável alvo não é conhecida na base de dados e a mineração categoriza apenas similaridade entre os dados (CAMILO e SILVA, 2009).

Nos sub-capítulos a seguir será expostos os principais conceitos de algumas técnicas de mineração, porém não serão aprofundados pois apenas a última é objeto de estudo deste trabalho.

2.3.1. Árvores de Classificação

Braga (2005, p 48) conceitua este técnica como:

> “Um procedimento hierárquico para predizer a classe de um objeto com base em suas variáveis preditoras [...]. A variável alvo é obviamente categórica e o método permite sua predição em função doa níveis observados das variáveis preditoras, as quais também devem ser categóricas. Caso não sejam é necessário codificá-las convenientemente.”

Este método requer uma amostra de treinamento, ou seja, uma parte dos registros da base de dados. Com base nesta amostra é possível construir com esta técnica uma árvore hierárquica onde cada nó representa uma variável da base de dados, como mostrado na Figura 4. Esta árvore será utilizada para classificar novos registros de acordo com o comportamento dos dados analisados pela amostra de treinamento.

Figura 4 - Exemplo de Árvore de Classificação (Fonte: SALVADOR et. al., 2009).

2.3.2. Regressão Logística

A regressão é uma técnica para valores quantitativos que objetiva estimar o valor que uma variável irá assumir baseado nos valores das demais variáveis, utilizando Aprendizado Supervisionado (CAMILO e SILVA, 2009).

A Regressão Logística (Equação 1) calcula um peso β para cada campo, de acordo com sua influência sobre a variável que se quer estimar. Os valores β_1 até β_r são utilizados na fórmula abaixo, onde p é a probabilidade da variável alvo que vai de 0 à 1, os valores de X_1 até X_r os valores dos demais campos na base de dados, e r o número de campos.

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \ldots + \beta_r X_r)}} \quad (1)$$

Os valores de p calculados por esta equação sempre ficam contidos no intervalo de 0 à 1, sendo que quanto mais o resultado se aproxima de 1, aumenta a probabilidade da variável em estudo ser positiva e quanto mais se aproxima de 0, amplia a possibilidade da variável ser negativa, formando uma curva como mostrado na Figura 5 (BRAGA, 2005).

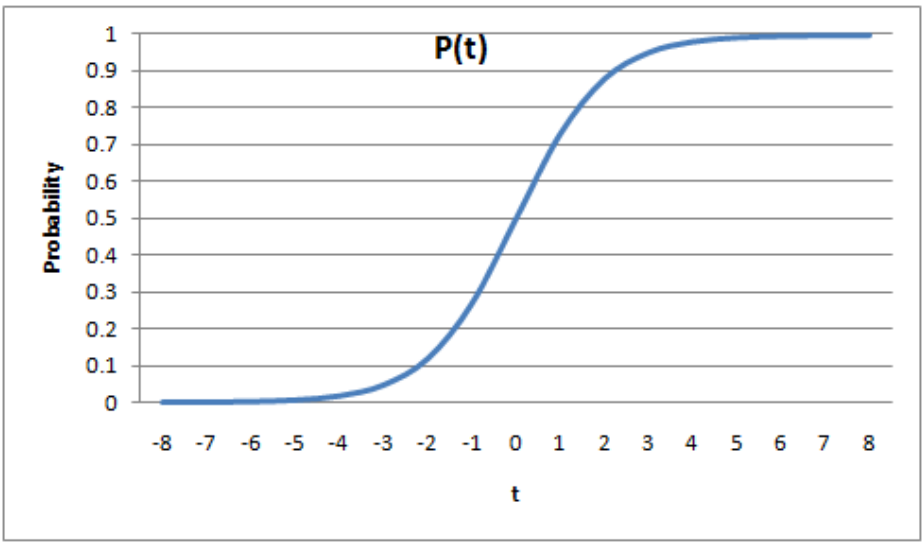


Figura 5 - Gráfico de Regressão Logística (Fonte: COLLINS, 2013).

2.3.3. Associação

As regras de associação são muito utilizadas na mineração de dados, e tem como objetivo avaliar a relação estre ddiferentes conjuntos de atributos. Desta forma pode-se descobrir, por exemplo, que é comum aos clientes que compram o produto X também comprar o produto Y. Uma informação como esta pode ser utilizada para alocar tais produtos de maneira que facilite a compra e , conseguentemente, maximize o lucro da empresa (SCHONHORST, 2010).

2.3.4. Clusterização

A Clusterização, ou Agrupamento é um método de Aprendizado Não-Supervisionado, que constrói grupos de registros, como mostrado na Figura 6, reunindo dados que apresentam similaridades entre si e se diferem das características de outros grupos. Este método não trabalha fazendo probabilidades ou classificações, pois seu objetivo consiste em analisar as semelhanças entre os atributos.

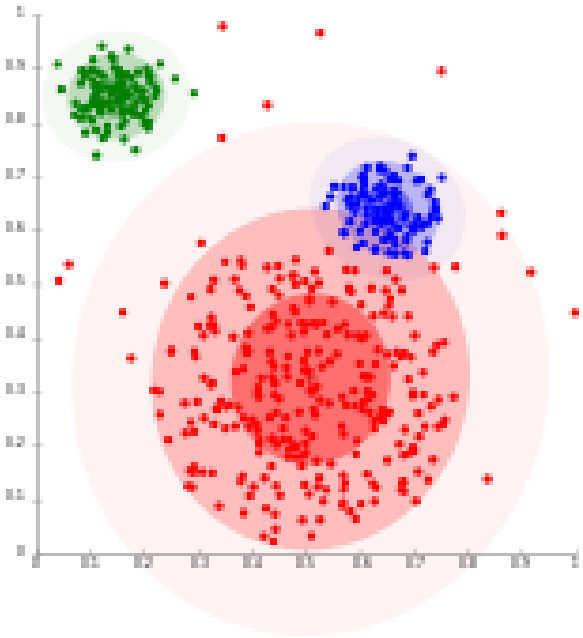


Figura 6 - Exemplo Clusterização (Fonte: WIKIPÉDIA, 2013).

Com base na função da técnica de clusterização, esta adequou-se ao objetivo deste trabalho e, portanto, será o método utilizado no desenvolvimento do modelo de Mineração de Dados descrito no Capítulo 3.

3. DESENVOLVIMENTO

Este capítulo descreve a preparação dos dados e a aplicação do algoritmo de mineração de dados proposta.

3.1. DESCRIÇÃO DO PROBLEMA

Atualmente o avanço tecnológico tem atingido os mais diversos tipos de estabelecimentos e são implementados *softwares* para gerenciar os mais diferentes tipos de informações. Esta tecnologia tem atingido também a área da saúde com a existência de diversos *softwares*, tanto para hospitais privados quanto para instituições públicas mantida pelo governo. Desta forma aumentou-se o número de informações produzidas sobre as sub-áreas da saúde, como, por exemplo, gerenciamento de internações, exames, nascimentos e mortalidade. O SIM (Sistema de Informações de Mortalidade) é um dos *softwares* criado pelo governo para gerenciar dados da saúde, especificamente sobre mortalidade infantil e produz uma enorme quantidade de dados. Entretanto, por ser um número tão grande de dados, é difícil interpretá-los, mas se fossem melhor compreendidos poderiam vir a produzir novos conhecimentos que facilitariam a compreensão das causas e fatores que influênciam a mortalidade infantil no Brasil e poderiam ser planejados projetos com o objetivo de minimizar tal mau em nosso país.

3.2. PREPARAÇÃO DOS DADOS

Antes da aplicação do algortimo de mineração, foi necessário estudar a base de dados utilizada neste trabalho com o intuito de melhor compreendê-la para que ao fim da aplicação pudesse ser feita a análise dos resultados. Foram estudados os atributos da base de dados e feita uma análise estatística dos campos que aparentaram ser de maior importância.

3.2.1. Sistema Único de Saúde

O Sistema Único de Saúde (SUS), mantido por impostos pagos pela população, foi criado em 1988 pela Constituição Federal para gerenciar a saúde no país de maneira que qualquer cidadão, independente de sua condição financeira, tenha acesso a atendimento público, tendo direito a tratamentos, internações e exames médicos.

Além dos postos de saúde e hospitais públicos, há alguns centros privados de saúde que fazem parte do SUS através de convênios e contratos quando o setor público do local não é suficiente para atender a população.

3.2.2. Base de Dados

Neste trabalho foi utilizada a base de dados do SIM, disponibilizadas para uso público pelo DATASUS[1], que é o Departamento de Informática do SUS.

O SIM é um sistema que gerencia as informações sobre mortalidade no Brasil criado em 1975 e informatizados em 1979, e armazena dados quantitativos e qualitativos regulares das declarações de óbito geralmente emitida pelo médico responsável. A partir de tais dados é possível realizar analises e avaliações para, desta forma, planejar medidas que possam reduzir a mortalidade por causas preveníveis ou evitáveis e otimizar a gestão da saúde no país.

A declaração de óbito, fonte de informação do SIM, é impressa em três vias e preenchida pela unidade de saúde onde ocorreu a comprovação do óbito. Essas vias são encaminhadas às Secretárias Municipais de Saúde para serem digitadas e inseridas no SIM local. Posteriormente, esses dados são encaminhados pela internet para as unidades estaduais, e em seguida para as federais. Essa fragmentação é de suma importância pois desta forma é possível realizar análises estatísticas mais específicas filtrando os dados por estado, ou região.

[1] http://www2.datasus.gov.br/DATASUS/index.php

O Ministério da Saúde disponibiliza publicamente os dados dos óbitos ocorridos entre 1979 e 2011 para serem utilizados como instrumento de pesquisa tanto pela internet, através do site do DATASUS, quanto em CDs enviados às Secretarias de Saúde e algumas Universidades.

Em 2005 a OMS (Organização Mundial da Saúde) avaliou os sistemas que gerenciam os dados de mortalidade, classificando sua qualidade como alta, média ou baixa. Nesta avaliação o SIM foi categorizado como qualidade intermediária e, desde então, o Brasil tem investido na potencialização deste sistema.

3.2.3. Pré-processamento da Base de Dados

Realizou-se um pré-processamento no banco de dados do SIM, onde se notou a existência de campos não preenchidos, como Código do Bairro, Escolaridade, e Ocupação. Além destes há alguns campos também não preenchidos que não constam na descrição dos dados cedida pelo DATASUS, como os campos OBITOGRA, CAUSABAS, e LINHAA. Para que estes valores faltantes não prejudiquem o resultado final da mineração, estes campos foram retirados da base de dados. Outros campos, como Idade da mãe e Quantidade de Filhos Vivos apresentavam uma pequena quantidade de valores armazenados como “99”, indicando dado não informado. Estes dados foram substituídos por “?”, que o weka interpreta como dado faltante, para que não fosse tratado como um valor numérico e influenciasse a mineração.

Notou-se também, entre os diversos atributos, o registro de informações que podem ser relevantes para a descoberta de conhecimento a respeito da mortalidade infantil, como o tipo do óbito (fetal ou não fetal), idade, sexo, raça, idade da mãe, quantidade de filhos vivos, quantidade de filhos mortos, gravidez (única, dupla, tripla ou mais), tempo de gestação e tipo do parto. Espera-se portanto que estes dados sejam suficientes para concretizar o estudo deste trabalho e alcançar seu objetivo.

3.2.4. Análises Estatísticas

Antes da implementação do algoritmo foi realizada uma análise estatística dos principais atributos da base de dados para melhor compreensão das informações. Nesta análise, apesar de 0.5% dos sexos não terem sido registrados, foi constatado que mais da metade dos óbitos infantis são do sexo masculino (55.6%), conforme mostrado no gráfico da Figura 7.

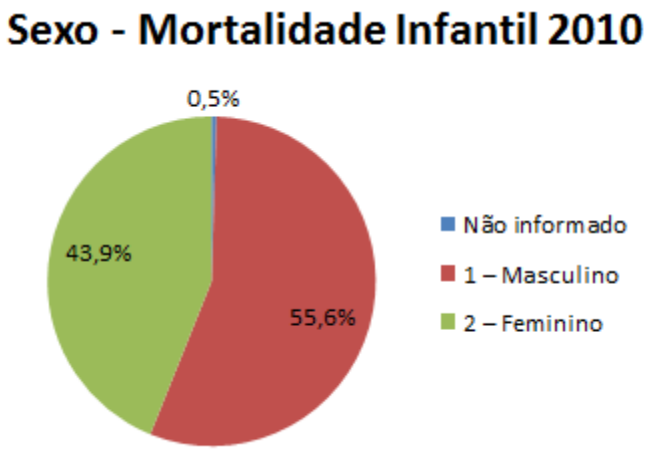


Figura 7 - Gráfico de Sexo da Mortalidade Infantil em 2010 (Fonte: Elaboração própria).

Foi constatado também que a média aritmética das idades da mãe foi de 25 anos, com moda, que é o valor que mais aparece na base de dados, de 21 anos. Entretanto o desvio médio, que é a média de quanto os dados diferem da média aritmética, foi de 5,85 anos, ou seja, dada a média de 25 anos, as idades das mães registradas nesta base de dados estão concentradas aproximadamente entre 20 e 30 anos (SPIEGEL, 1987).

Também foram analisados os dados a respeito da raça dos óbitos maternos, mostrados na Figura 8, onde constatou-se que a grande maioria das raças dos óbitos se dividem entre branca e parda, com 16457 (41,3%) e 17282 registros (43,3%), respectivamente, somando 33739 registros (84,6%). Contudo é possível afirmar que a quantidade de óbitos das raças branca e parda superam notavelmente o número de óbitos das demais raças nesta base de dados.

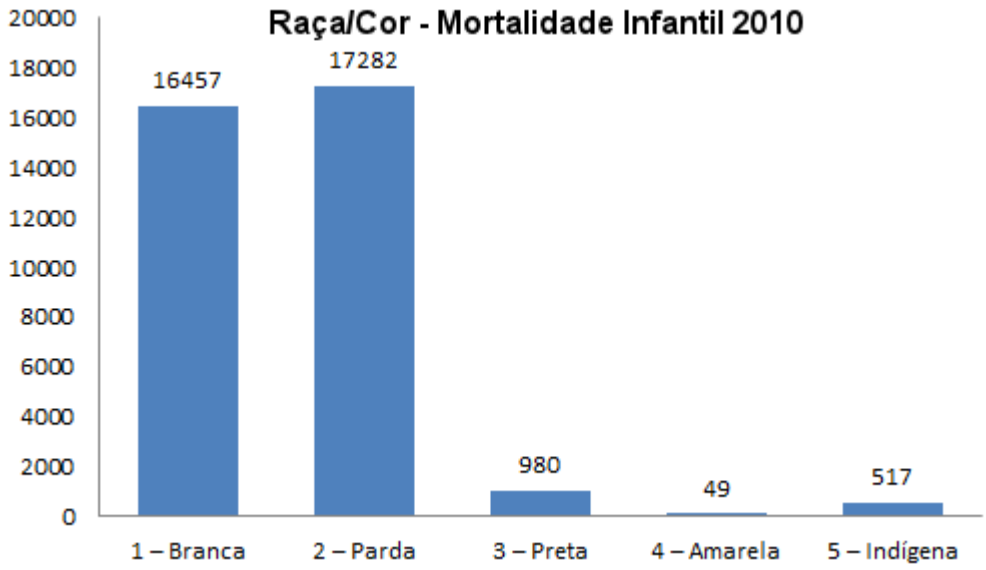


Figura 8 - Gráfico de Raça e Cor da Mortalidade Infantil em 2010 (Fonte: Elaboração própria).

Além disso, foram analisadas as informações da quantidade de filhos mortos que as mães tiveram antes do óbito registrado e foi encontrada a média aritmética de 2,02 filhos, e a moda de 0 filhos com 21651 registros, ou seja, a maioria dos óbitos registrados são de primogênitos.

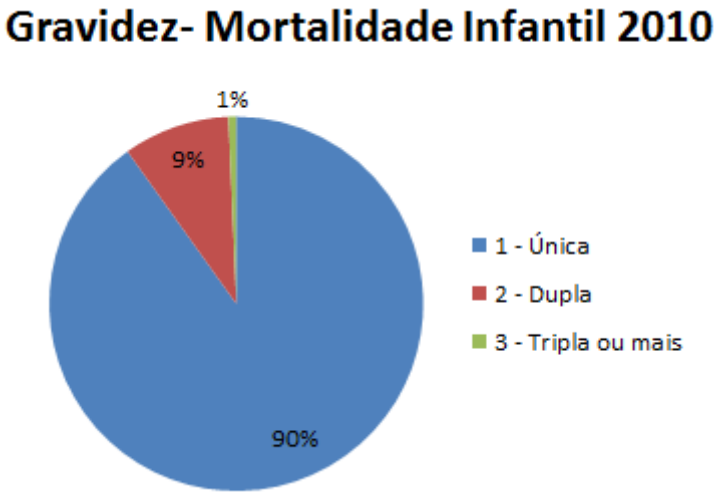


Figura 9 - Gráfico de Gravidez da Mortalidade Infantil em 2010 (Fonte: Elaboração própria).

Em seguida foi construído um gráfico com os dados do tipo de gravidez (Figura 9), onde verifica-se que a evidente maioria das gravidez foram únicas, o que é esperado, uma vez que este tipo de gravidez é a mais comum. Além do tipo de gravidez, foi analisado o tipo de parto, conforme mostrado na Figura 10, onde observa-se que 19306 (56%) óbitos infantis tiveram parto vaginal e 15119 (44%) tiveram cesária, desta forma, considerando o total de registros, a desigualdade entre os dois tipos de parto (12%) pode ser considerada pequena.

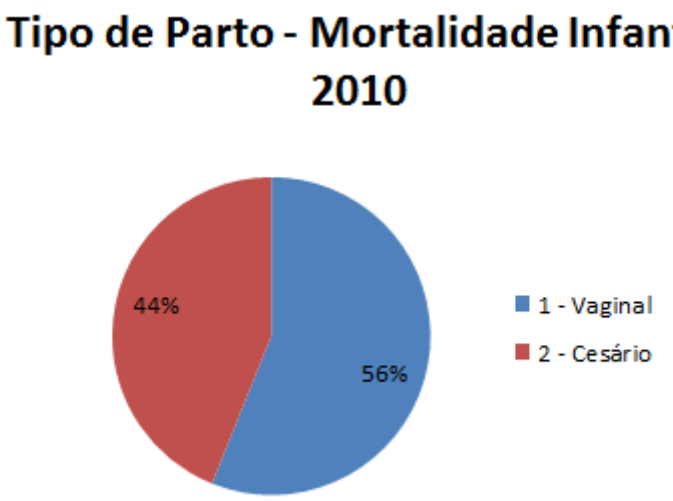


Figura 10 - Gráfico de Tipo de Parto da Mortalidade Infantil em 2010 (Fonte: Elaboração própria).

3.2.5. Transformação dos Dados

A Base de Dados disponibilizada no site do DATASUS se encontra no formato .dbc, que é a compactação de arquivos .dbf. Para leitura destes dados fez-se necessário realizar o *download* do sotware Tabwin, disponibilizado para livre acesso pelo Ministério da Saúde no mesmo site. Com este *sotware* foi possível descompactar o arquivo e, após a descompactação, lê-lo utilizando o *software* LibreOffice Calc e, assim, visualizar os dados. Feito isto, foi salvo um arquivo com a extensão .csv, que é um arquivo onde as informações da planilha é convertida para formato texto. A partir deste arquivo foi gerado pelo *software* Weka um arquivo .arff que consiste também em um arquivo texto padrão deste *software*, com um

cabeçalho onde são descritos os atributos especificando o tipo de dado de cada um, seguidos pelos dados em si. Este será o arquivo utilizada para implementação do algoritmo de mineração de dados no Weka.

3.3. WEKA

O Weka, acrograma de *Waikato Environment for Knowledge Analysis*, é um *software* de mineração de dados desenvolvido pela Universidade de Waikato, Nova Zelândia, e licenciado com a *General Public License*, que é uma licença que permite que o código fonte seja acessado e estudado por qualquer usuário.

O Weka foi desenvolvido em Java e, portanto, uma vez configurada a JVM (*Java Virtual Machine* - máquina virtual Java) é possível executá-lo na maioria dos sistemas operacionais. Possui uma *interface* gráfica para acessar a base de dados, e apresentar resultados como gráficos e tabelas, e disponibiliza também uma API (*Application Programming Interface*) que pode ser incorporada ao desenvolvimento de outras aplicações para realizar funcionalidade de mineração de dados.

A tela principal do Weka permite ao usuário escolher entre os quatro tipos de aplicações disponíveis, sendo eles *Explorer*: Ambiente onde é possível explorar os dados, implementando algoritmos de mineração, e visualisar gráficos específicos de cada atributo e gráficos da relação entre os mesmos; *Experimenter*: Ambiente onde se faz testes e avaliações estatísticas; *KnowledgeFlow*: Ambiente com função equivalente ao *Explorer*, porém neste a exploração dos dados é feita através de um fluxograma criado pelo usuário com os módulos disponíveis; e *Simple CLI*: Interface para mineração de dados em linha de comando.

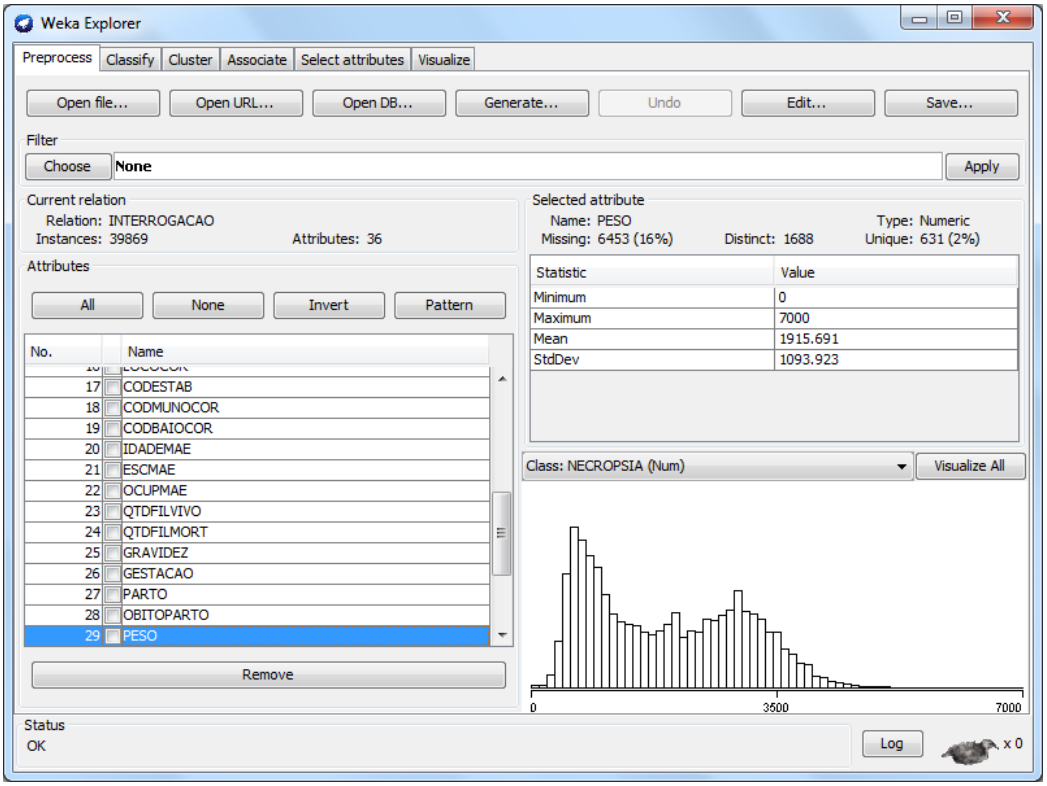


Figura 11 – Ambiente Explorer (Fonte: Elaboração própria).

Ao abrir o ambiente *Explorer* (Figura 11) são disponibilizados na parte superior os seguintes recursos:

- *Preprocess*: Ambiente onde pode ser feito um pré-processamento dos dados, aplicar filtros, e visualizar informações específicas de cada atributo como maior e menor valor e média;
- *Classify*: Ambiente que permite treinamento e testes de algoritmo de classificação;
- *Cluster*: Ambiente para implementação de agrupamentos de dados, também chamado de clusterização; e
- *Visualize*: Ambiente onde é montado um esquema 2D dos dados, evidenciando a relação de cada um dos atributos com os demais (Figura 12).

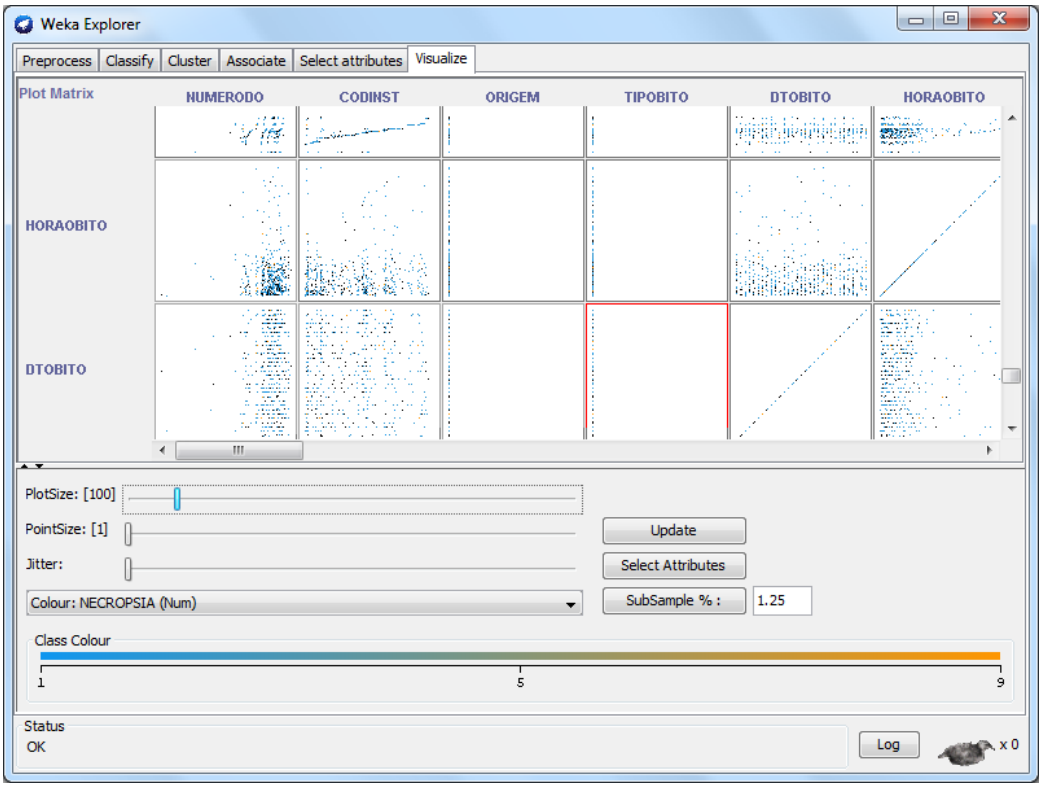


Figura 12 – Ambiente *Visualize* com esquemas 2D dos dados (Fonte: Elaboração própria).

Na parte inferior do ambiente *Emplorer* encontra-se um botão chamado *Log*, que ao ser clicado abre um histórico com todas as ações que foram realizadas durante o uso atual do Weka.

3.4. ALGORITMO *EXPECTATION-MAXIMIZATION (EM)*

Para uso neste trabalho foi selecionado o algoritmo *Expectation Maximization* (EM), pois este é indicado por Ng e McLachlan (2009) para clusterização em base de dados com dados incompletos, por possuir convergência global confiável e estabilidade numérica. O *EM* estima os valores desconhecidos e os parâmetros, e repete esta estimação até que os erros se estabilizem.

Inicialmente, o *EM* calcula uma nova matriz de dados (B) com dados estimados de acordo com os valores conhecidos na base de dados, interpretada também como uma matriz (A), e a partir destas duas matrizes é construída uma

terceira matriz (C) os dados conhecidos de A e os valores faltantes preenchidos pelos estimados em B, assim como ilustrado na Figura 13, onde a cor clara representa os valores conhecidos e a cor escura os valores estimados.

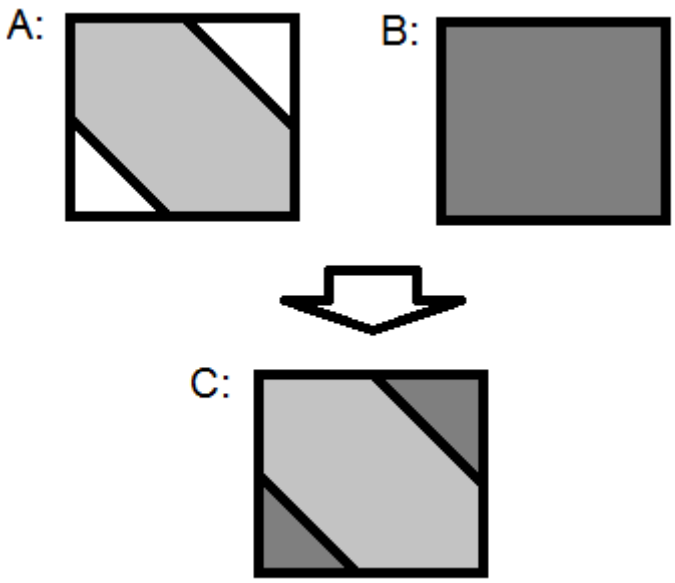


Figura 13 – Ilustração do passo *Expectation* (Fonte: Elaboração própria).

Desta forma, a construção da matriz C pode ser definida na Equação (2), em que M representa uma matriz de máscara, onde $M_{ij} = 0$ se o elemento ij da matriz A for desconhecido e 1 caso seja conhecido, ⊙ é o produto ponto-a-ponto, ou seja, a multiplicação de cada elemento ij de uma matriz com o respectivo elemento ij da segunda matriz, e $[1]$ representa uma matriz de mesma ordem que as demais com todos seus elementos preenchidos por 1.

$$C = A \odot M + B \odot ([1] - M) \tag{2}$$

Tendo encontrado C, considera-se h a hipotese atual referente ao conjunto de parâmetros θ e h' a hipótese revisada a cada iteração do *EM*.

Em seguinda o *EM* estima os parâmetros θ atraves da busca pela hipótese h' que maximiza a Equação (3), onde $P(C|h')$ representa a verossimilhança

dos dados completos C na hipotese h'. Além disto, introduz-se o valor esperado da Equação (3), pois como B é gerado por uma distribuição de probabilidade, esta é uma variável aleatória, tornando sua derivada C tambem uma variavel aleatoria (LUNA, 2004).

$$E[\ln P(C|h')] \tag{3}$$

Para encontrar a hipotese que maximiza esta expressão, é necessario estimar a distribuição de probabilidades de C, porém esta distibuição é desconhecida pois é determinada utilizando os parâmetros θ. Por este motivo, o *EM* utiliza a hipótese atual para estimar a distribuição de probabilidades de C. Esta transformação é demonstrada pela função Q, conforme Equação (4), supondo que o conjunto de parâmetros θ seja igual à hipotese atual h.

$$Q(h'|h) = E[\ln P(C|h')|h, A] \tag{4}$$

Este algoritmo se dividide, portanto, nos dois seguintes passos:

1. *E* (*Expectation*): onde é calculado $Q(h'|h)$ com a hipotese atual h e os dados A para determinar a distribuição de probabilidade sobre C utilizando a Equação (4);

2. *M* (*Maximization*): onde troca-se a hipotese h pela h', maximizando a função Q, conforme a Equação (5).

$$h = \arg\max Q(h'|h)' \tag{5}$$

3.5. EXPERIMENTOS E RESULTADOS

Após carregar o arquivo .arff da base de dados no ambiente *Explorer* do Weka, foi aplicado o filtro *StratifiedRemoveFolds* para retirar da base uma parcela de dados que será utilizada para teste do algoritmo no momento da aplicação. Desta

forma, dividiu-se a base de dados em duas, uma com 3987 registros, ou seja, 10% da base original, e outra com 35882 registros. A partir da base de dados maior foi feita uma análise inicial com o Visualize, e em seguida foi aplicado o algoritmo *Expectation Maximization* usando a base menor como conjunto de teste.

3.5.1. Análise com o *Visualize*

A partir do ambiente *Visualize* foi realizada uma análise preliminar da relação entre os atributos da base de dados. Entretanto o número de atributos dificultou esta análise, pois os esquemas 2D montados pela ferramenta não permite ver claramente a relação de todos os dados.

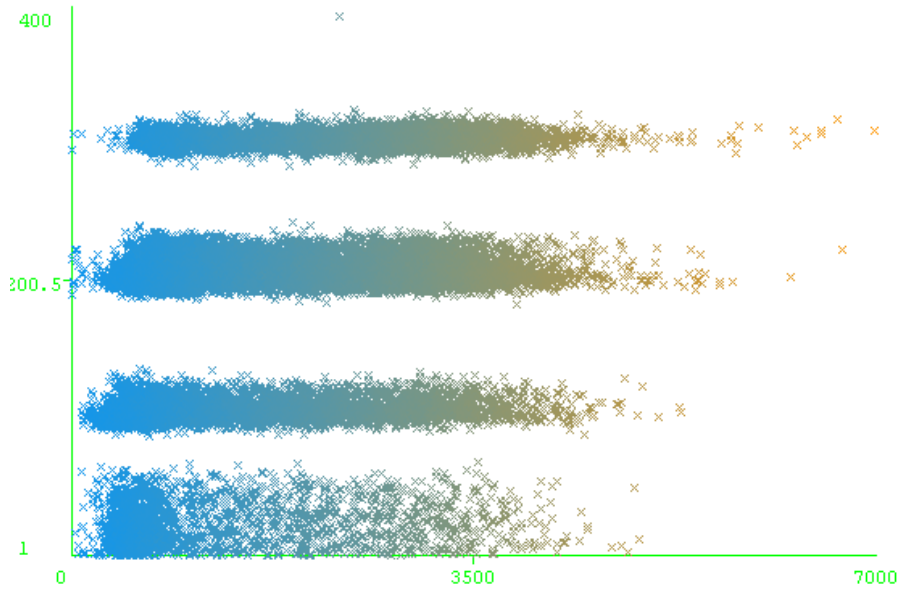
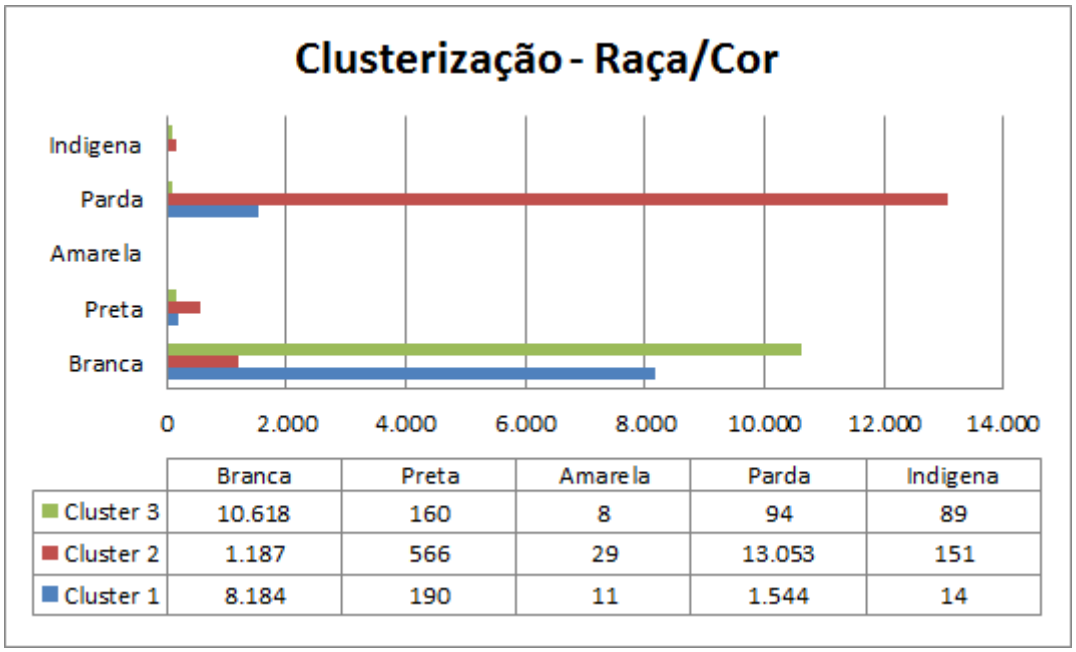


Figura 14 – Gráfico comparativo entre peso e idade (Fonte: Elaboração própria)

Ainda assim, notou-se uma tendência de agrupamento no gráfico da relação entre peso e idade, conforme mostrado na Figura 14, onde os dados são nitidamente separados em quatro grupos. Contudo esta é uma informação já esperada, pois, considerando que são dados de crianças de até cinco anos, pressume-se que quanto maior for a idade conseguentemente maior será o peso.

3.5.2. Aplicação do Algoritmo *EM*

Foi aplicado o algoritmo *EM*, utilizando como conjunto de teste a base de dados construída pelo filtro *StratifiedRemoveFolds*. O algoritmo separou os dados em três grupos, chamados também de *clusters*, sendo o segundo *cluster* o maior, com 80% do registros. No primeiro *cluster* foram designados apenas 6% dos dados da base original, correspondendo ao menor dos três *clusters*. Notou-se que no menor *cluster* a média da idade da mãe é de 27,5 anos, sendo maior que a dos outros demais *clusters* (24,5 e 24,4 anos) e também maior que a média geral (25 anos).



	Branca	Preta	Amarela	Parda	Indigena
Cluster 3	10.618	160	8	94	89
Cluster 2	1.187	566	29	13.053	151
Cluster 1	8.184	190	11	1.544	14

Figura 15 – Clusterização do EM para Raça/Cor (Fonte: Elaboração própria)

Devido ao fato do segundo *cluster* ser maior que os demais, é esperado que englobe maior parte dos dados de cada atributo, fazendo com que o número de óbitos de cada raça seja maior neste cluster. Porém foi observado que no *cluster* 2 uma grande quantidade de óbitos pertencem à raça parda, superando excessivamente as demais raças, conforme mostrado na Figura 15. Isto é notável, pois um grande número de óbitos desta raça se concentra em um único *cluster*, ao contrário da raça branca, que apesar de também ter grandes números, se divide entre os *clusters* 1 e 3.

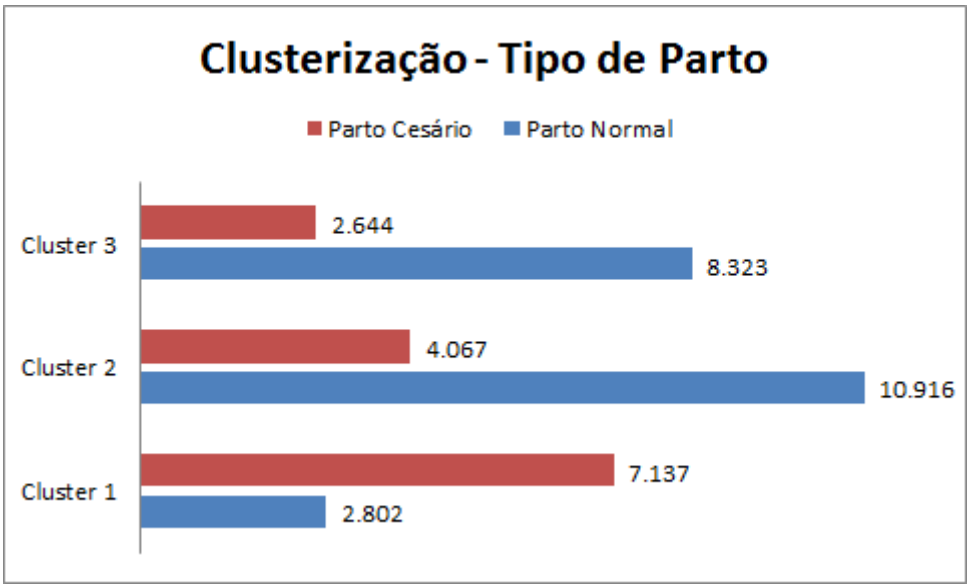
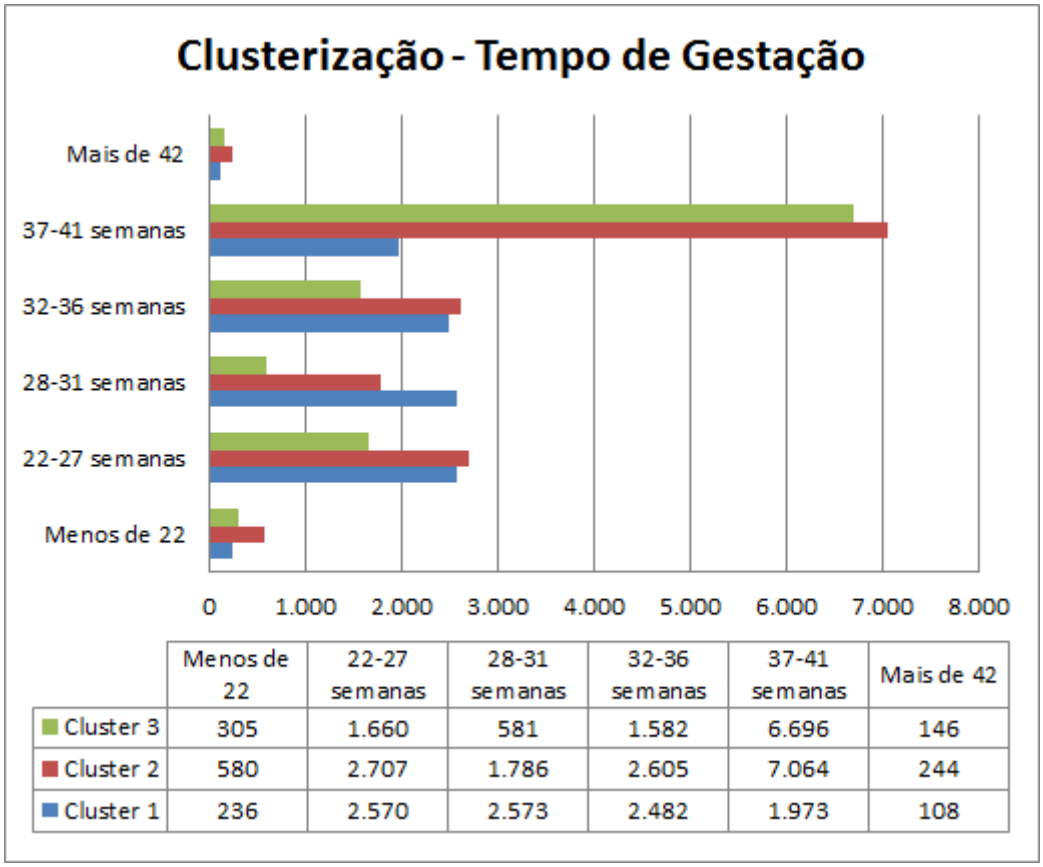


Figura 16 – Clusterização do EM para tipo de parto (Fonte: Elaboração própria)

Também foi constatado que nos dois maiores *clusters* o número de óbitos que sofreram parto normal superam em grande número os óbitos provenientes de parto cesário, com números que ultrapassam o dobro, conforme mostrado na Figura 16. Apenas no *cluster* 1, que contém 6% dos dados, o número de parto cesário é maior que o de parto normal. Além disto, observou-se, como mostrado na Figura 17, que grande parte dos óbitos fetais ocorrem no perído entre 37 e 41 semanas de gestação, com a maior quantidade estando no cluster 2, e um número proporcionalmente grande no *cluster* 3. Observa-se também que os óbitos relativos ao período entre 28 e 31 semanas concentram-se no cluster 1.

Todas estas informações podem ser úteis no planejamento de novos projetos de minização da taxa de mortalidade infantil, uma vez que ficou evidente que o grupo de óbitos infantis relativos à raça parda que sofrem parto normal, provavelmente por dependerem apenas do setor público, excede grandemente os demais. Entretanto, em um algoritmo de clusterização, os dados não são divididos tendo em vista apenas um atributo, mas sim com a influência e relação entre estes atributos. Por este motivo, a análise do *cluster* se mostra extremamente complicada,

uma vez que também não se deve observar cada atributo individualmente, mas sim os *clusters* como um todo e os dados pertencentes a cada um.



	Menos de 22	22-27 semanas	28-31 semanas	32-36 semanas	37-41 semanas	Mais de 42
Cluster 3	305	1.660	581	1.582	6.696	146
Cluster 2	580	2.707	1.786	2.605	7.064	244
Cluster 1	236	2.570	2.573	2.482	1.973	108

Figura 17 – Clusterização do EM para tempo de gestação (Fonte: Elaboração própria)

4. CONSIDERAÇÕES FINAIS

A taxa de mortalidade infantil representa quantas crianças menores de um ano morrem a cada mil crianças nascidas vivas. Essa taxa é um importante indicador da situação do país, pois está relacionado com as condições do sistema de saúde e do saneamento básico oferecido à população. Portanto, é de interesse comum que sejam tomadas providências para diminuir tal índice.

Com o avanço tecnológico, o setor da saúde tem produzido um grande número de informações a respeito dos indicadores de saúde, sendo um deles a mortalidade infantil. Porém, a base de dados produzida é muito extensa, o que torna dificil analisá-la. Para possibilitar esta análise foi feito uso da Mineração de Dados, que busca padrões, anomalias, similaridades ou discordância entre os dados para transforma-los em informações úteis. Com isto, esperava-se produzir conhecimento suficiente para ajudar na criação de futuros projetos de minimização da taxa de mortalidade infantil em nosso país.

Após expor os conceitos e os dados históricos a respeito da mortalidade infantil e da Mineração de Dados, foi apresentado o funcionamento da ferramenta Weka, utilizada na aplicação proposta neste trabalho. Foi selecionado o algoritmo *Expectation Maximization*, aplicando-o em uma base de dados cedida pelo Ministério da Saúde para uso público. Por meio desta aplicação, os dados foram divididos em três grupos com similariedades entre os atributos, tais como raça e tipo de parto. Entretanto, a hipotese confirmada neste trabalho foi a de que apesar de descobrir alguns padrões, o conhecimento resultante apenas da aplicação do algoritmo *Expectation Maximization* nesta base de dados é insuficiente para auxiliar a criação de novos projetos de redução da mortalidade infantil, podendo ser utilizado como apoio no planejamento destes projetos, mas não como única fonte de informação e certeza de sucesso.

Por este motivo, sugere-se para trabalhos futuros a aplicação de outros algoritmos de clusterização nesta mesma base de dados para comparar os resultados obtidos e avaliar o melhor entre eles. Sugere-se também a junção da base de dados de mortalidade infantil com os registros de todos os nascimentos no país para criação de um modelo que faça uso de um algoritmo de predição,

estimando a probabilidade de ocorrer óbito infantil dadas as novas informações a respeito dos pacientes.

5. REFERÊNCIAS

ANDRADE, M. M. de. **Introdução à Metodologia do Trabalho Científico**. 9º ed. São Paulo: Atlas, 2009.

BRAGA, L. V. **Introdução à Mineração de Dados**. 2º ed. Rio de Janeiro: E-papers. 2005.

CAMILO, C. O. SILVA, J. C. **Mineração de Dados**: Conceitos, Tarefas, Métodos e Ferramentas. Instituto de Informática Universidade Federal de Goiás. 2009.

COLLINS, S. E. Logistic Regression Demo. Disponível em: <https://catalyst.uw.edu/workspace/collinss/9542/57085>. Acesso em: 04 Dez. 2013. 15h30.

FAYYAD, U; PIATETSKY-SHAPIRO, G; SMYTH, P. **From Data Mining to Knowledge Discovery in Databases**. California: American Association for Artificial Intelligence, 1996.

FRANCISCO, W. de C. e. **Mortalidade Infantil no Brasil**. Disponível em: <http://www.brasilescola.com/brasil/mortalidade-infantil-no-brasil.htm>. Acesso em: 23 Out. 2013. 13h00.

FRANK, A., ASUNCION, A. **UCI Machine Learning Repository.** Irvine / CA. Disponível em: <http://archive.ics.uci.edu/ml>. Acesso em: 01 Mar. 2013. 14h30.

HALL, M., FRANK, E., HOLMES, G., PFAHRINGER, B., REUTEMANN, P., WITTEN, I. H. **The WEKA Data Mining Software**: An Update; SIGKDD Explorations, Volume 11, Issue 1. 2009.

IBGE. **Evolução e perspectivas da mortalidade infantil no Brasil**. Rio de Janeiro, 1999. 45 p. (Estudos e Pesquisas. Informação Demográfica e Socioeconômica, n. 2).

LUNA, E. O. **Algoritmos EM para Aprendizado de Redes Bayesianas a partir de Dados Incompletos**. 2004. 120 f. Tese (Mestrado em Ciência da Computação) – Universidade Federal de Mato Grosso do Sul, Campo Grande, MS. 2004.

NG, S. MCLACHLAN, G. EM. In: WU, X. Kumar, V. **The Top Ten Algorithms in Data Mining**. Boca Raton/FL: CRP Press. 2009. p 93 - 116

OLIVEIRA, I. C. **Aplicação de Data Mining na Busca de um Modelo de Prevenção da Mortalidade Infantil**. 2001. 104 f. Tese (Mestrado em Engenharia de Produção e Sistemas) – Universidade Federal de Santa Catarina, Florianópolis, S.C. 2001.

REZENDE, S. O., PUGLIESI, J. B., MELANDA, E. A., PAULA, M. F. de. Mineração de Dados. In: REZENDE, S. O. **Sistemas Inteligentes**: Fundamentos e Aplicações. Barueri/SP: Manole. 2003. p 307 - 335.

SALVADOR, H. G. et. al. **Vedalogic um método de Verificação de Dados Climatológicos Apoiado em Modelos Minerados**. Revista Brasileira de Meteorologia, São Paulo, vol.24, no.4, 2009.

SCHONHORST, G. B. **Mineração de Regras de Associação Aplicada à Modelagem dos Dados Transacionais de um Supermercado**. 2010. 80f. Tese (Mestrado em Ciências em Engenharia de Produção) – Universidade Federal de Irajubá, Itajubá MG. 2010.

SPIEGEL, M. R. **Estatística**. 2° ed. São Paulo: McGraw-Hill do Brasil. 1985.

WHO. **World Health Statistics 2013**. Genebra/Suíça: World Health Organization Press. 2013

WIKIPÉDIA. **Cluster Analysis.** Disponível em: <http://en.wikipedia.org/wiki/Cluster_analysis>. Acesso em: 24 Out. 2013. 13h45.

WILLIAMS, G. **Data Mining with Rattla and R**: The Ar of Excavating Data for Knowledge Discovery. New York/NY: Springer. 2011. p 3 - 5.

WITTEN, I. H., FRANK, E. **Data Mining**: Pratical Machine Learning Tools and Techniques. 2º ed. San Francisco/CA: Morgan Kaufmann. 2005.